INTELIGENCIA DE NEGOCIOS ISIS3304

"Proyecto 1 Analítica de textos"

Profesor: Fabian Peña

Grupo 3

Daniela Uribe – d.uribe1

Valentina Calderón - v.calderonm

Santiago Forero – s.forerog2

Índice

1.Entendimiento del negocio y enfoque analítico

6.7 Objetivos y criterios de éxito desde el punto de vista del negocio

El objetivo central del proyecto es desarrollar un modelo de clasificación basado en aprendizaje automático para vincular textos con los Objetivos de Desarrollo Sostenible (ODS) de las Naciones Unidas (ONU). Esto permitirá analizar la información recopilada por Fondo de Poblaciones de las Naciones Unidas (UNFPA) durante procesos de planificación territorial, comprender las opiniones de las comunidades locales sobre los ODS y facilitar la toma de decisiones en políticas públicas.

1.2 ¿Qué es un ODS?

Un ODS, o "Objetivo de Desarrollo Sostenible," es un conjunto de metas globales adoptadas por la Organización de las Naciones Unidas (ONU). La Agenda 2030 para el desarrollo sostenible, adoptada por la ONU, se basa en 17 Objetivos de Desarrollo Sostenible (ODS) y 169 metas. En este proyecto se trabajarán los ODS 3, 4 y 5.

Referencia: https://ods.dnp.gov.co/es/objetivos/salud-y-bienestar

1.3 ODS del proyecto e impacto

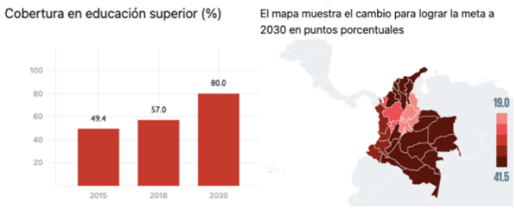
El ODS 3 tiene como objetivo primordial garantizar una vida saludable y promover el bienestar para todas las edades. Esto implica reducir la mortalidad infantil y materna, así como mejorar el acceso a servicios de salud de alta calidad. En Colombia, este objetivo cobra una importancia significativa, ya que según datos del DANE, la tasa de mortalidad materna por cada 100 mil nacidos vivos fue de 53.7 en 2015, disminuyendo a 51 en 2018, con la meta gubernamental nacional de reducirla a 32 para 2030. Es crucial mencionar que existen disparidades regionales en el país, siendo las regiones más alejadas, como el Caribe, el Pacífico, la Amazonia y la Orinoquía, las que enfrentan mayores desafíos en términos de salud. Para abordar esta problemática, se implementó el Programa de Acción Integral en Hospitales "Ai Hospital" en zonas estratégicas, que busca mejorar la infraestructura, dotación y sistemas de información hospitalaria para brindar servicios de salud más efectivos. (DNP)

Metas Regionales a 2030

Fuente: Cálculos DNP. DANE - Estadísticas Vitales (EEVV).

El ODS 4 se enfoca en garantizar una educación inclusiva, equitativa y de alta calidad para todos, promoviendo oportunidades de aprendizaje a lo largo de toda la vida. En Colombia, es de gran relevancia ya que, según el DANE, en 2015 menos del 50% de la población tenía acceso a educación superior, y la meta nacional es alcanzar el 80% para 2030. Sin embargo, se observan desigualdades regionales, con un acceso más limitado en áreas menos centralizadas, las cuales requieren de mayor esfuerzo para lograr cambios. Para abordar este desafío, el gobierno implementó el programa "GENERACIÓN E", que beneficia a jóvenes en situación de vulnerabilidad económica y les brinda acceso a programas de alta calidad a través de los componentes de Equidad y Excelencia, con la proyección de beneficiar a un gran número de jóvenes en el futuro. (DNP)

Metas Regionales a 2030



Fuente: Cálculos DNP. DANE - Estadísticas Vitales (EEVV).

Por último, el ODS 5 busca lograr la igualdad entre géneros y empoderar a todas las mujeres y niñas. En Colombia, según el DANE, en 2015, el porcentaje de mujeres en cargos directivos del Estado colombiano era del 43.5%, con la meta nacional de llegar al 50% para 2030. Aunque las diferencias por departamento no son tan marcadas, las regiones del Pacífico y el Caribe muestran un menor progreso en este aspecto. (DNP)

Metas Regionales a 2030

Fuente: Cálculos DNP. DANE - Estadísticas Vitales (EEVV).

Analizar estos tres ODS (3, 4 y 5) son fundamentales en Colombia debido a los desafíos que enfrenta el país en áreas de salud, educación y equidad de género. El enfoque en estas metas contribuirá al desarrollo sostenible y al bienestar de la población colombiana.

Referencia: https://ods.dnp.gov.co/es/objetivos/salud-y-bienestar

1.4 Contacto estudiantes de estadística

Correo: i.beltranc@uniandes.edu.co Fecha en la que nos reuniremos: 14 de octubre del 2023

Canal: Reunión por Zoom, compartir documentos por Drive

1.5 Tabla

Oportunidad/ problema Negocio	El problema principal identificado en este contexto es la necesidad de clasificar eficientemente grandes cantidades de información textual recopilada por el Fondo de Poblaciones de las Naciones Unidas (UNFPA) en relación con los Objetivos de Desarrollo Sostenible (ODS) de la Agenda 2030. Actualmente, esta tarea de clasificación de textos se realiza manualmente y requiere un gran esfuerzo humano y recursos. El desafío radica en la automatización de este proceso para acelerar y mejorar la precisión de la clasificación de textos según los ODS.
Enfoque analítico	Tipo de aprendizaje: supervisado Tarea de aprendizaje: Clasificación Técnicas: Arboles de decisión - Random Forest Classifier Algoritos: BoW, TF-IDF , RandomForest
Organización y rol dentro de ella que se beneficia con la oportunidad definida	El principal beneficiario de esta oportunidad sería el Fondo de Poblaciones de las Naciones Unidas (UNFPA), que podría utilizar el modelo de clasificación automática de textos para agilizar y mejorar su capacidad de relacionar la información recopilada con los ODS. Esto permitiría una evaluación más eficiente de las políticas públicas y el impacto social en relación con los ODS.

Contacto con experto externo al proyecto	La organización que se beneficia con la solución es el Fondo de Población

2. Entendimiento y preparación de los datos

2.1 Perfilamiento y análisis de la calidad de los datos

Para cada dato se proporcionan 2 columnas: 'Textos_espanol' y 'sdg', siendo la primera la información textual recopilada de tipo ‘object’ al ser texto y la segunda un número entre 3, 4 y 5 de tipo ‘int64’ que simboliza el objetivo de desarrollo sostenible al que está asociado. Se evidenció que en total hay 3000 datos que se dividen de forma pareja entre los objetivos de desarrollo sostenible 3, 4 y 5. Es decir, que 1000 datos tienen un valor de ‘sdg’ de 3, otros 1000 de 4 y otros 1000 de 5.

Posteriormente se notó que no todos los documentos estaban en español, ya que 9 de los 3000 datos estaban en un idioma diferente como inglés o francés. Se decidió no eliminarlos puesto que, aunque estén en otros idiomas, al probar el modelo eliminando estos datos se aumentaba el error, probablemente porque algunas palabras claves se conservan en otros idiomas. Así que se conservaron finalmente las palabras. Además, se eliminaron los acentos y caracteres especiales de los documentos con el fin de garantizar una mejor precisión al elegir palabras del idioma español utilizando un encoder de latín.

A. Completitud: Después de analizar todos los datos se pudo comprobar que no había valores nulos, de modo que se garantiza completitud.

B. Unicidad: No hay ningún valor que se espere que sea único al no haber ningún id, de modo que no existe un problema de unicidad.

C. Consistencia: No hay inconsistencias ya que el tipo de los datos coincide con lo esperado al ser tipo ‘object’ las cadenas de texto e ‘int64’ los datos numéricos.

D. Validez: No se hallaron problemas de validez por datos puesto que los datos de 'Textos_espanol' no tiene valores atípicos y 'sdg' tampoco, al estar siempre entre 3, 4 y 5.

Por lo que, al no haber ningún problema de completitud, unicidad, consistencia ni validez, no hay que arreglar rangos, formatos o categorías ni eliminar datos atípicos, columnas o valores nulos. Más allá de transformar los tipos ‘object’ en ‘str’.

2.2 Tratamiento de los datos

Para el tratamiento de los datos previo al dominio, las técnicas y los algoritmos seleccionados para resolver las tareas se dividieron los datos entre los valores de entrenamiento (train) y de prueba (test), esto en un porcentaje de 70% y 30% respectivamente. Posteriormente, se realizó la vectorización del texto con el fin de poder ser analizado, para lo cual inicialmente se creó una lista llamada stop_words que contiene palabras comunes en español que generalmente no aportan mucho significado en el análisis de texto. Estas palabras se utilizarán más adelante para filtrarlas y eliminarlas del texto, ya que no aportan información valiosa en muchos casos.

Después, para la preparación del primer algoritmo se usó "Bag of Words", un modelo que se encarga de contar la frecuencia de las palabras en el texto. Se configuró un modelo de vectorización llamado ‘bow’, al cual se le proporcionó información como el ‘tokenizer’, las palabras comunes a eliminar (las que se definieron en stop_words) y se convirtió todo el texto a minúsculas. A continuación, se aplicó el modelo a los textos en español del conjunto de datos de entrenamiento previamente definido con el fin de crear una representación numérica de los textos en español.

De forma similar, para la preparación del segundo algoritmo se usó “TF-IDF”, primero se configuró otro modelo de vectorización llamado ‘tfidf’, al cual también se le especificó el tokenizador, las palabras comunes a eliminar y la conversión a minúsculas. Este modelo se aplicó a los mismos textos en español del conjunto de datos de entrenamiento creando una representación numérica diferente de los textos en español. Una vez realizadas ambas transformaciones se comparó la cantidad de palabras únicas en los textos de entrenamiento después de aplicar cada vectorización y se vio que en ambos casos fue 15439. De forma que ambos métodos de vectorización tienen el mismo tamaño de vocabulario, tal como debería pasar.

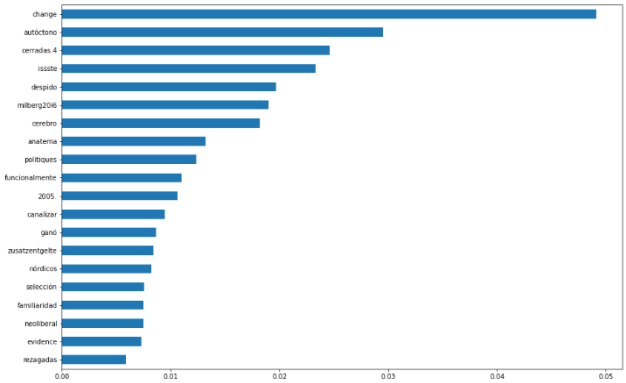
3. Modelado y evaluación

3.1 Algoritmo 1: BoW + RandomForest

Como primer algoritmo primero se usó Bag of Words (BoW), que convierte texto en datos numéricos. Para su funcionamiento inicialmente, cada texto se descompone en palabras individuales y se crea un "saco" que contiene todas las palabras únicas presentes en el texto, sin importar su orden o contexto, de las cuales se filtran las ‘stop_words’ como se explicó en el paso anterior. Cada palabra se representa como un vector independiente, y la frecuencia de aparición de cada palabra se registra en un conjunto de datos. Esto permite convertir el texto en datos numéricos que se pueden utilizar en tareas como clasificación de documentos o en un análisis de sentimientos. Sin embargo, BoW no conserva información sobre la estructura gramatical o el significado de las palabras, ya que trata a cada palabra por separado, por lo que el contexto gramatical muchas veces se pierde y puede llegar a ser una limitación cuando se requiere un análisis más profundo del texto. Posterior a esto se aplica el algoritmo de “Random Forest” que toma estos números y trata de predecir categorías o etiquetas para ese texto. Este algoritmo busca diferentes formas de hacer este proceso, probando diferentes configuraciones, y elige la mejor configuración que funcione bien en función de cómo acierta y cómo se equivoca en la tarea de predicción.

Se seleccionó este modelo debido a su simplicidad y eficiencia. BoW es fácil de implementar, computacionalmente rápido y flexible en términos de dominio e idioma, por lo que es adecuado en este contexto donde se necesitan procesar grandes cantidades de texto. Aunque BoW no tiene en cuenta el significado asociado al contexto de las palabras, su capacidad para contar la frecuencia de las palabras lo hace valioso en este contexto donde la semántica detallada no es algo crítico. Respecto a Random Forest, este es un algoritmo de clasificación flexible que suele funcionar bien en una variedad de contextos, como en este caso con las características numéricas generadas por BoW.

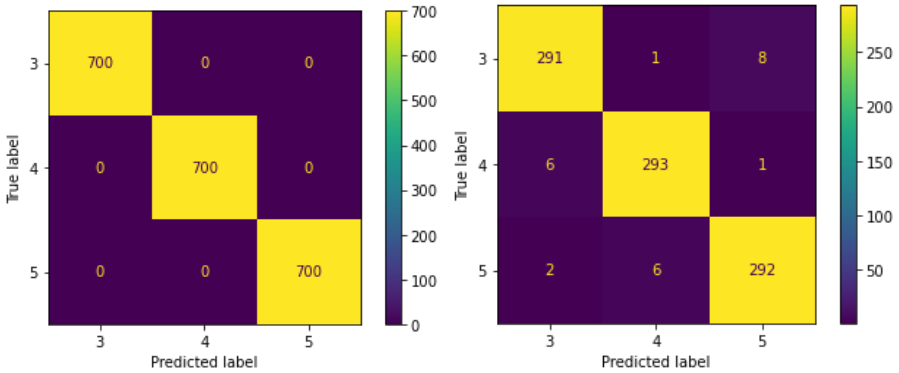
A continuación, se muestran las 20 palabras más influyentes en la clasificación de los textos según este modelo.



A continuación, se muestran las 10 palabras que más influyen junto con los temas de los textos relacionados a cada uno de ellos y las ocurrencias totales.

Palabra	Temas relacionados	Ocurrencias totales
Change	Salud, género	4
Autóctono	Educación	1
Cerradas	Educación, género	3
Issste	Salud	3
Despido	Salud, educación	4
Cerebro	Salud, educación	19
Anatema	Género	1
Politiques	Educación	1
Funcionalmente	Salud	1
Canalizar	Género	2

Es razonable que la mayor parte de las palabras estén relacionadas con temas de salud, educación y género ya que los objetivos 3,4,5 tratan de salud y bienestar, educación de calidad e igualdad de género.



Se puede ver que la predicción se realizó a la perfección para el conjunto de train. Para el conjunto de prueba, en su mayoría se realizan bien las predicciones. El de mejor predicción es el del objetivo 4, luego el 5 y por último el 3. El error más alto está en textos del objetivo 3 que se clasifican como objetivo 5. Es posible dado que en la calidad de salud y bienestar también hay desigualdad de género, por lo que hay una mezcla de ambos temas. Asimismo, los siguientes errores más altos son textos del objetivo 5 que están haciendo clasificados por igual del objetivo 4 y de tipo 4 que se predicen de tipo 3. Esto se da ya que el género también es un factor nuevamente determinante de la salud y de la educación, por lo que puede haber muchos textos que toquen ambos temas y sea difícil decidir que pertenecen más al 5 que a los otros dos. Asimismo, se podría decir que los objetivos 4 (educación de calidad) y 5 (igualdad de género) son determinantes en la calidad de vida y en el bienestar. Por lo tanto, es razonable que muchos textos del objetivo 4 se clasifiquen en el objetivo 3, pues esta trata de bienestar.

Dado que se tiene más de una clase, se hallarán las métricas en parámetro micro para calcular las métricas globalmente contando el total de TP, FN y FP en todas las clases.

Train: Precision: 1.0 - Recall: 1.0 - F1: 1.0

Test: Precision: 0.9733 - Recall: 0.9733 - F1: 0.9733

Como se puede ver las métricas para los datos de entrenamiento son perfectas pues todas las predicciones estuvieron acorde a las clases de los datos como se observó anteriormente. En cuando a las de los datos de prueba, las métricas son muy buenas y todas dan un valor igual, por lo que se podría decir que los errores de falsos positivos y falsos negativos están balanceados.

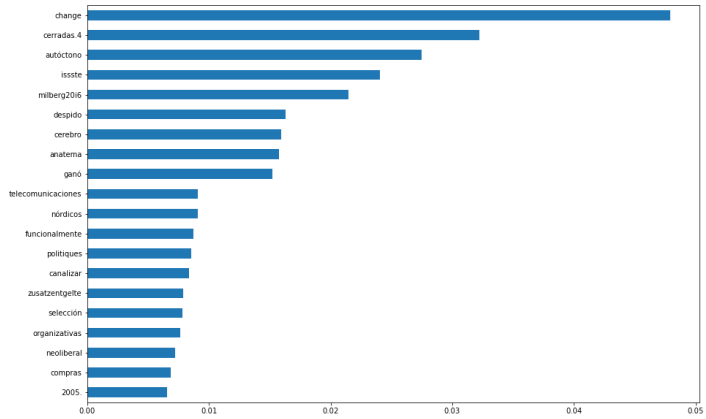
Integrante que trabajó en este punto: Daniela Uribe

3.2 Algoritmo 2: TF-IDF + RandomForest

Como segundo algoritmo se seleccionó Term Frequency - Inverse Document Frequency, (TF-IDF), que convierte texto en datos numéricos según la importancia de una palabra en un documento dentro de un conjunto de documentos, como en este contexto. Para esto, recibe lista de palabras definida en la preparación de los datos (después de eliminar las stop-words) y calcula dos valores diferentes: la frecuencia de término (TF), que mide cuántas veces aparece una palabra en un documento específico, y la frecuencia inversa del documento (IDF), que mide que tan común es esa palabra en todos los documentos. Multiplicando estos dos valores, TF-IDF asigna un peso a cada palabra en un documento, lo que permite identificar palabras clave o términos significativos. Posterior a esto se aplicó nuevamente el algoritmo de “Random Forest” que toma estos números y trata de predecir categorías o etiquetas para ese texto como se explicó previamente.

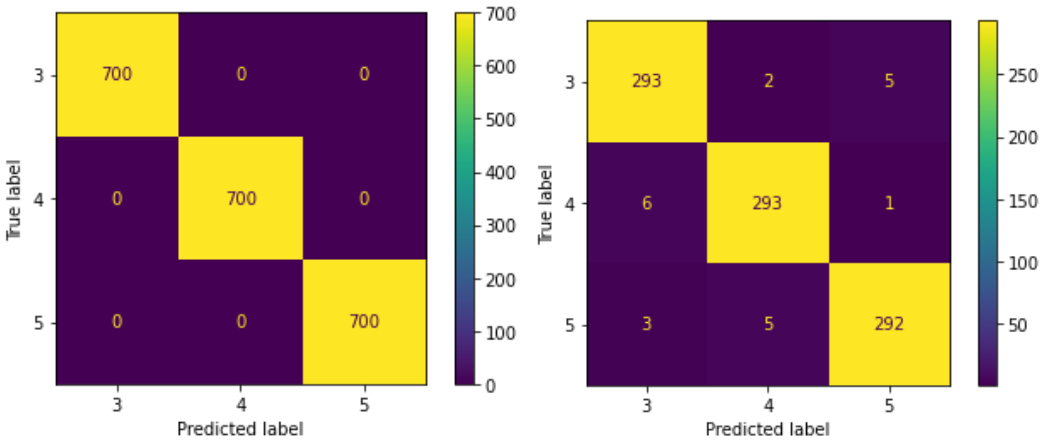
Se seleccionó este modelo debido a su capacidad para resaltar las palabras específicas y poco comunes en un documento en comparación con un conjunto más amplio de documentos. Esto resulta importante en este análisis de texto donde se busca identificar la información relevante y clasificar los documentos de manera efectiva de acuerdo con los objetivos dados. Aunque al igual que BoW, TF-IDF no tiene en cuenta el significado asociado al contexto de las palabras, su capacidad para contar la frecuencia de las palabras lo hace valioso en este contexto donde la semántica detallada no es algo crítico. Respecto a Random Forest, este es un algoritmo de clasificación flexible que suele funcionar bien en una variedad de contextos, como en este caso con las características numéricas generadas por TF-IDF.

A continuación, se muestran las 20 palabras más influyentes en la clasificación de los textos según este modelo.



A continuación, se muestran las 10 palabras que más influyen junto con los temas de los textos relacionados a cada uno de ellos y las ocurrencias totales

Palabra	Temas relacionados	Ocurrencias totales
Change	Salud, género	4
Autóctono	Educación	1
Cerradas	Educación, género	3
Issste	Salud	3
Despido	Salud, educación	4
Cerebro	Salud, educación	19
Anatema	Género	1
Ganó	Salud, educación, género	3
Telecomunicaciones	Salud	2
Nórdicos	Salud, educación, género	13



Se puede ver que la predicción se realizó a la perfección para el conjunto de train. Con respecto al modelo anterior se mantiene la cantidad de errores para textos del objetivo 4 clasificados como del 3 y textos del objetivo 4 clasificados como del objetivo 5. Disminuye la cantidad de error del objetivo 3 y se mantiene constante el error en el 4 y en el 5. Estos cambios se pueden dar dado que ahora con la vectorización TDF-IDF se tiene en cuenta la cantidad de veces que una palabra aparece en diferentes textos, por lo tanto, parece ser que hay ciertas palabras relacionadas a la salud que aparecen en tantos textos del objetivo 5, que se genera esta peor clasificación. Asimismo, hay tantas palabras de educación que aparecen en textos del objetivo 3, que se genera esta peor clasificación.

Train: Precision: 1.0 - Recall: 1.0 - F1: 1.0 Test: Precision: 0.9755 - Recall: 0.9755 - F1: 0.9755

Como se puede ver las métricas para los datos de entrenamiento son perfectas pues todas las predicciones estuvieron acorde a las clases de los datos como se observó anteriormente. En cuando a las de los datos de test, las métricas son muy buenas y todas dan un valor igual, por lo que se podría decir que los errores de falsos positivos y falsos negativos están balanceados. Asimismo, las métricas son mejores a las del modelo anterior por lo que se concluye que en cuanto a métricas este modelo es mejor.

Integrante que trabajó en este punto: Valentina Calderón

3.3 Algoritmo 3: (Bow/ TDFIF) + RandomForest + RandomizedCV

Como tercer y último algoritmo se seleccionó una combinación de dos componentes: el primero siendo “Bag of Words” que convierte texto en valores numéricos dependiendo de cuántas veces aparece cada palabra en el texto, y el segundo siendo “Random Forest” que toma estos números y trata de predecir categorías o etiquetas para ese texto como se explicó previamente. La diferencia con el primer algoritmo es que esta vez también se usó “RandomizedCV” que es una técnica que busca automáticamente las mejores configuraciones para el modelo “Random Forest” variando con diferentes hiperparámetros. Esto se hace probando múltiples combinaciones hasta encontrar la configuración más precisa y eficiente del modelo. Todo esto por medio de un Pipeline.

La elección de este modelo, que combina Bag of Words (BoW) con Random Forest, se justifica por varias razones. En primer lugar, BoW permite capturar la frecuencia de las palabras en un texto, lo que puede ser informativo para tareas de clasificación de texto como ya se explicó previamente junto a las ventajas que trae. Respecto a Random Forest, este es un algoritmo de clasificación flexible que suele funcionar bien en una variedad de contextos, como en este caso con las características numéricas generadas por BoW. Respecto a RandomizedCV, se utilizó ya que la búsqueda aleatoria de hiperparámetros permite explorar diferentes configuraciones y encontrar la mejor para maximizar el rendimiento del modelo en métricas específicas, como el F1-score. Por lo que, en conclusión, esta combinación ofrece un enfoque fuerte y adaptable para tareas de

clasificación de texto con la posibilidad de mejorar el rendimiento mediante la optimización de hiperparámetros.

Train: Precision: 1.0 - Recall: 1.0 - F1: 1.0 Test: Precisión: 0.977 - Recall: 0.977 - F1: 0.977

Como se puede ver las métricas para los datos de entrenamiento son perfectas al igual que en los modelos anteriores pues todas las predicciones estuvieron acorde a las clases de los datos. En cuando a las de los datos de prueba, las métricas son buenas y todas dan un valor igual, por lo que se podría decir que los errores de falsos positivos y falsos negativos están balanceados. Igualmente, las métricas son un poco mejores que las de los dos modelos anteriores por lo que se concluye que este modelo es el mejor. Es importante resaltar que en los casos de prueba que realizó este algoritmo se realizó con la vectorización tanto de TD IDF como de BoW.

Integrante que trabajó en este punto: Santiago Forero

3.4 Algoritmo seleccionado

Al comparar las métricas de calidad de los tres algoritmos se analiza que como los modelos con RandomForest iniciales que no incluyen el RandomizedCV tienen peores métricas, el mejor modelo es esté último, el que se escogió finalmente.

4. Resultados

Resultado en términos “comunes” para que la organización los entienda

Después de realizar un análisis completo de la calidad de los datos, se han aplicado tres algoritmos de clasificación, y se ha seleccionado el modelo que incluye el RandomizedCV como el más adecuado para la tarea de clasificación de textos relacionados con los Objetivos de Desarrollo Sostenible. Los resultados son prometedores, con métricas de precisión, exhaustividad y valor F1 en niveles muy altos, lo que indica un buen rendimiento en la clasificación de textos.

Análisis de métricas de calidad:

Algoritmo 1:
Train: Precision: 1.0 - Recall: 1.0 - F1: 1.0 Test: Precision: 0.9733 - Recall: 0.9733 - F1: 0.9733
Algoritmo 2:
Train: Precision: 1.0 - Recall: 1.0 - F1: 1.0 Test: Precision: 0.9755 - Recall: 0.9755 - F1: 0.9755
Algoritmo 3:
Train: Precision: 1.0 - Recall: 1.0 - F1: 1.0 Test: Precision: 0.977 - Recall: 0.977 - F1: 0.977

En resumen, los tres algoritmos muestran un rendimiento sobresaliente en el conjunto de entrenamiento, con valores perfectos de precisión, exhaustividad y F1. Lo anterior se afirma porque: La precisión es igual a 1 lo que indica que el modelo acierta en todas las predicciones positivas. Que el Recall sea igual a 1 indica que el modelo encuentra todos los casos positivos en el conjunto de datos de entrenamiento. Que F1 sea igual a 1, indica que la media armónica entre precisión y recall es perfecta. En el conjunto de prueba, los algoritmos 1 y 2 tienen un rendimiento muy similar con valores de alrededor del 97.33% y 97.55%, mientras que el algoritmo 3 muestra un rendimiento ligeramente superior, con valores alrededor del 97.77%. Esto sugiere que los algoritmos son efectivos en la clasificación de datos en ambos conjuntos, pero el Algoritmo 3 destaca especialmente en el conjunto de prueba.

Cómo aportan a la consecución de los objetivos: La alta calidad de las métricas sugiere que las predicciones derivadas del modelo serán altamente **precisas**. Un modelo de clasificación preciso es fundamental para el proyecto de apoyo a UNFPA y para relacionar textos con los Objetivos de Desarrollo Sostenible (ODS) de manera adecuada. Esto facilita la clasificación eficiente de textos

según los ODS, así como la interpretación y el análisis de datos textuales recopilados a través de diversas fuentes en procesos de planificación participativa. En este orden de ideas, ellos pueden utilizar nuestro modelo (el **Pipeline automatizado**) para insertar otros conjuntos de datos y que genere una clasificación automática.

Estrategias/Sugerencias que puede implementar la organización: En general, la clasificación de las opiniones les permitirá conocer de mejor forma a las comunidades locales evaluadas, construyendo estrategias específicas y personalizadas que se adapten a las necesidades y prioridades de cada grupo. Esto optimiza la eficacia de sus acciones y el impacto de sus programas de desarrollo sostenible. Algunas de las estrategias y decisiones que se ven facilitadas incluyen la personalización de programas de desarrollo, la focalización de recursos, la colaboración con actores locales y la facilidad en la toma de decisiones en políticas públicas

Sugerencias: Como sugerencia, se recomienda ejercer precaución debido a la similitud de los objetivos, ya que todos están orientados hacia el desarrollo sostenible. Esto es importante porque algunas combinaciones de objetivos pueden resultar en clasificaciones erróneas.

Relevancia: La información obtenida es importante para la organización ya que les permite identificar los Objetivos de Desarrollo Sostenible (ODS) a los cuales las personas otorgan una mayor prioridad. Con esta comprensión, la organización puede adaptar sus estrategias y enfoques para abordar de manera más efectiva las necesidades y preocupaciones prioritarias de las comunidades locales, lo que contribuye a un enfoque más centrado y eficiente en la promoción de los ODS.

5. Mapa de actores

Rol dentro de la empresa	Tipo de actor	Beneficio	Riesgo
Fondo de Poblaciones de las Naciones Unidas (UNFPA)	Usuario-cliente	Facilita la evaluación y seguimiento de políticas públicas relacionadas con los ODS, permitiendo una toma de decisiones más informada y basada en datos, lo que mejora la eficacia de sus proyectos y actividades	Si el modelo de clasificación no funciona correctamente, podría llevar a interpretaciones erróneas de la información recopilada y decisiones inadecuadas basadas en esos datos
Entidades públicas y gubernamentales	Financiador	Obtienen información más rápida y precisa sobre cómo sus políticas impactan en la sociedad en relación con los ODS, lo que puede ayudar en la toma de decisiones más efectivas.	En caso de que el modelo no funcione es dinero mal invertido que pudo usarse mejor. Además, depender demasiado del modelo automatizado podría no capturar matices importantes o tener sesgos, lo que podría llevar a decisiones incorrectas.
Universidad de los Andes (entidad colaboradora)	Proveedor	Contribuye al desarrollo de tecnología y soluciones que tienen un impacto significativo en la mejora de políticas públicas y la calidad de vida de las comunidades locales.	Responsabilidad en la calidad y ética de los modelos de clasificación, ya que pueden afectar decisiones importantes en políticas públicas.
Habitantes locales	Beneficiado	Sus opiniones y preocupaciones se ven reflejadas en las políticas y acciones de desarrollo sostenible, lo que	Vulneración de la privacidad y seguridad de los datos personales si la información se maneja de forma inadecuada.

		potencialmente podría mejorar su calidad de vida.	

6. Trabajo en equipo

6.1 Roles

Líder del proyecto: Daniela Uribe Líder del negocio: Santiago Forero

Líder de datos: Valentina Calderón Líder de analítica: Daniela Uribe

6.2 Tiempos

- Lanzamiento y planeación de trabajo en equipo: 3 horas
- Entendimiento del negocio y enfoque analítico: 1 hora
- Mapa de actores: 30 minutos
- Entendimiento y preparación de los datos: 2 horas y media
- Desarrollo primer algoritmo: 2 horas
- Desarrollo segundo algoritmo: 1 hora y media
- Desarrollo tercer algoritmo: 2 horas
- Análisis de resultados y creación del video: 2 horas y media
- Revisión grupal: 3 horas y media
- Cierre: 1 hora

6.3 Retos y soluciones

- Calidad de los datos: Garantizar que los datos utilizados para entrenar el modelo sean precisos y representativos ya que los datos de baja calidad pueden afectar negativamente la calidad de las predicciones. Para esto se realizó un análisis cuidadoso de los datos dados para hallar errores y solucionarlos.
- Calidad de los modelos: Garantizar que los modelos generados sean precisos y cumplieran con los objetivos del negocio. Se revisaron los errores asociados de los datos de entrenamiento y prueba para comprobar esto.
- Correcto análisis de los resultados: Saber cómo analizar los resultados, ya que a primera vista no es evidente sacar conclusiones, de modo que se tuvo que ser muy cuidadoso y leer reiteradas veces el enunciado para asegurarse que el análisis hecho coincidiera con la realidad de los modelos.

6.4 Repartición 100 puntos

- Daniela Uribe: 34 puntos Valentina Calderón: 33 puntos Santiago Forero: 33 puntos

6.5 Cómo mejorar para la siguiente entrega:

- Apoyarnos más en profesores y monitores para algunas preguntas difíciles
- Darle más prioridad al enunciado desde el inicio para tenerlo en cuenta durante todo el desarrollo del proyecto
- Comunicarnos activamente entre nosotros más allá de las reuniones en casos de necesidad

6.6 Reuniones:

Reunión lanzamiento y planeación – Realizada el 26 de septiembre

Reunión de ideación – Realizada el 26 de septiembre

Reuniones de seguimiento – Realizada el 3 de octubre

Reunión de finalización – Realizada el 9 de octubre